future internet

MDPI

Article

Automated Business Goal Extraction from E-mail Repositories to Bootstrap Business Understanding

Marco Spruit [1,2,3,*], Marcin Kais [3] and Vincent Menger [3]

[1] Leiden University Medical Center (LUMC), Campus The Hague, Leiden University, Turfmarkt 99, 2511 DC The Hague, The Netherlands
[2] Leiden Institute of Advanced Computer Science (LIACS), Leiden University, Niels Bohrweg 1, 2333 CA Leiden, The Netherlands
[3] Department of Information and Computing Sciences, Utrecht University, Princetonplein 5, 3584 CC Utrecht, The Netherlands; marcinkais@gmail.com (M.K.); v.j.menger-2@umcutrecht.nl (V.M.)
* Correspondence: m.r.spruit@lumc.nl

Abstract: The Cross-Industry Standard Process for Data Mining (CRISP-DM), despite being the most popular data mining process for more than two decades, is known to leave those organizations lacking operational data mining experience puzzled and unable to start their data mining projects. This is especially apparent in the first phase of Business Understanding, at the conclusion of which, the data mining goals of the project at hand should be specified, which arguably requires at least a conceptual understanding of the knowledge discovery process. We propose to bridge this knowledge gap from a Data Science perspective by applying Natural Language Processing techniques (NLP) to the organizations' e-mail exchange repositories to extract explicitly stated business goals from the conversations, thus bootstrapping the Business Understanding phase of CRISP-DM. Our NLP-Automated Method for Business Understanding (NAMBU) generates a list of business goals which can subsequently be used for further specification of data mining goals. The validation of the results on the basis of comparison to the results of manual business goal extraction from the Enron corpus demonstrates the usefulness of our NAMBU method when applied to large datasets.

Keywords: CRISP-DM; business understanding; e-mail data understanding; business goal generation; natural language processing; knowledge discovery

Citation: Spruit, M.; Kais, M.; Menger, V. Automated Business Goal Extraction from E-mail Repositories to Bootstrap Business Understanding. *Future Internet* **2021**, *13*, 243. https://doi.org/10.3390/fi13100243

Academic Editor: Carlos Filipe Da Silva Portela

Received: 29 July 2021
Accepted: 18 September 2021
Published: 23 September 2021

1. Introduction

The emergence of Data Mining (DM) as a specific domain led to the creation of high-level models, structuring the data mining processes. CRISP-DM, being the most popular method and the de facto worldwide standard for data mining [1,2] provides a roadmap for DM projects as illustrated in Figure 1, by specifying their individual phases—Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment—and going deeper into particular tasks. It is however widely known that CRISP-DM does not provide sufficient support for the first step of the data mining process, namely the understanding of the problem owner's (i.e., business insider's) concerns [3]. This step, which is referred to as the Business Understanding phase within CRISP-DM, is considered to be critical, to the point where it can make or break the entire project [4]. There is a particular contrast of the perceived importance of business understanding to its actual implementation, which is often conducted in an ad-hoc manner [5].

Sharma et al. [6] explain this phenomenon by describing the general lack of support towards how the implementation of this phase should be performed. Ref. [7] conclude that, even though CRISP-DM is an improvement when compared to the previous standards in the data mining community, it is still not mature enough to deal with the complex problems it needs to address, with the apparent lack of business modelling procedures, formal tools, or methods for the business understanding phase. As a result, there is much

room for improvement when it comes to structuring the initial stages of the data mining process. This also explains the many task-specific extensions to the generic knowledge discovery process, which have been proposed over the years, including those in the domains of interactive data mining [8], big data processing [9], reproducible research [10] and personalised recommendation systems [11]. In this paper, we propose the NAMBU method for automatic business goal extraction from an organization's textual resources. This method can be embedded within the business understanding phase of CRISP-DM, and thus serve as a bootstrap for the current standards.

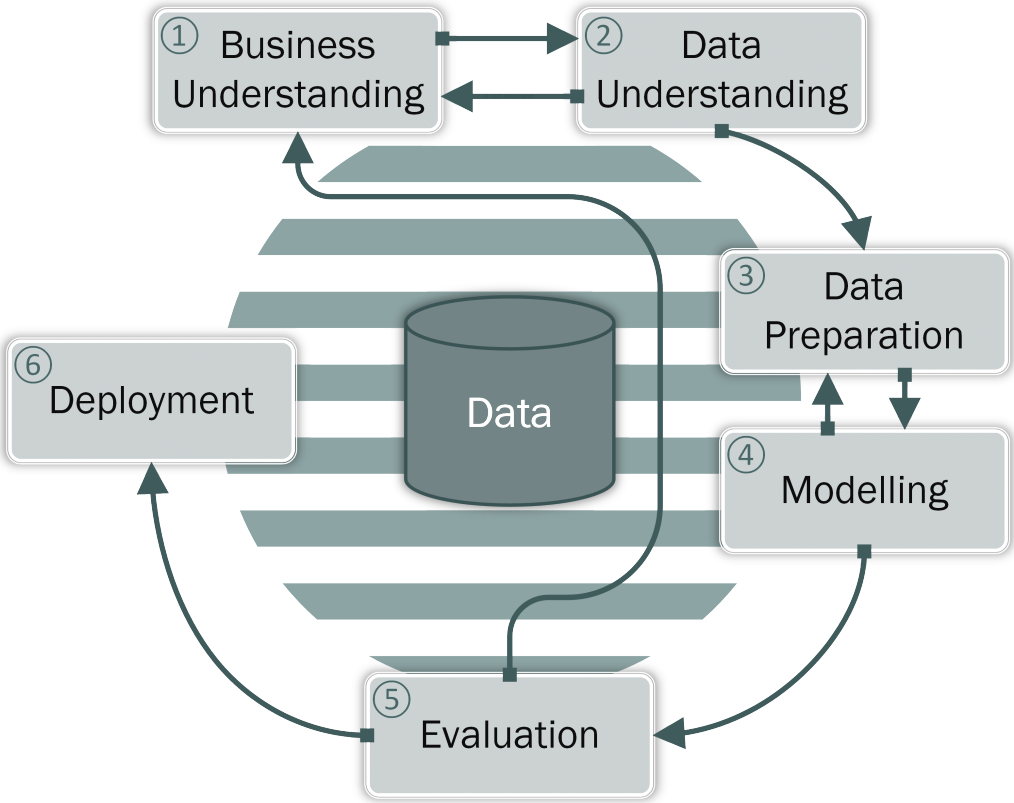

Figure 1. The CRISP-DM process starts with the Business Understanding phase, but does not specify how to determine business goals.

Natural Language Processing (NLP) is an area of research and application of a theoretically motivated range of computational techniques for analyzing, understanding, and manipulating naturally occurring texts at one or more levels of linguistic analysis for the purpose of performing and achieving desired tasks with human-like language processing accuracy [12]. Information extraction is a domain, where natural language processing methods are applied to identify and isolate relevant information from unstructured textual data. In our case, business goals are the target information.

As for the choice of sources, organizational e-mail exchange repositories are considered to be a valuable source of expertise, with an easy-to-mine set of communications between people in the social network [13]. These semi-anonymous communications, through media such as e-mail, also referred to as "weak ties" in businesses, are crucial to the flow of knowledge within organizations [14]. The general knowledge management strategy of an organization can be supported by capturing the employees' communication records and subjecting them to further analysis when needed [15].

The objective of this research was to develop the Meta-Algorithmic Model (MAM) shown as a recipe for using these textual resources to identify business goals explicitly stated in communications between the members of the organizations, to facilitate the business understanding phase of CRISP-DM [16,17]. Thus, the scientific contribution of

this research is threefold. First, we present the NLP-Automated Method for Business Understanding (NAMBU) as a reusable meta-algorithmic model for other researchers. Second, we extract business goal phrases automatically from free texts, such as emails, by employing a novel combination of goal-driven information extraction and rule-based NLP techniques. Third, we describe its empirical results using a real-world email dataset processed by a working software prototype implementation.

The remainder of this paper is structured as follows. Section 2 of this research introduces the developed NAMBU method, and explains its main constituents. Section 3 presents the results and performance measures of their evaluation. Section 4 discusses the limitations of the research, and proposes possible areas for further improvements. Section 5 sums up the conclusions and contributions of this work.

2. Methods

The NAMBU business goal extraction method was developed by further building upon previous approaches to automatic goal identification found in scientific literature. The supporting artefact was developed in the Java programming language, with the use of Stanford CoreNLP Natural Language Processing Toolkit, which is an annotation pipeline framework, providing most of the common core natural language processing steps [18]. To develop the tool, a sample set of textual corporate data was needed. Due to privacy concerns and legal restrictions, sets like that are quite hard to obtain. For this reason, the so-called Enron Corpus was used [19]. Enron Corporation was an American energy company, based in Houston, which bankrupted on 2 December 2001 in the wake of its massive accounting fraud scandal. The dataset of 619,446 e-mails sent and received by 158 Enron employees was collected in 2002 during the investigation into the company's collapse, as commissioned by the Federal Energy Regulatory Commission. It was later purchased and released to researchers, being the first publicly available mass collection of corporate e-mails. The process of business goal extraction along our method is divided into five steps: Extract sentences, Parse syntax, Identify goals, Extract (and format) goals, and finally, Filter results. These steps are explained in the following subsections of this document and presented as a Process-Deliverable Diagram (PDD) in Figure 2, a UML-derived modeling technique which integrates UML Activity and UML Concept diagram representations [20].

2.1. Extract Sentences

To minimize the computational load on the natural language processing tasks, especially when it comes to such an abundant resource as the corporate e-mail repository, only the sentences containing goal-related keywords, thus suspected to contain business goals, are subjected to NLP analysis. As suggested by [21–23], a list of goal-related keywords was compiled using their contributions, and further expanded to account for various forms of the keywords (e.g., improve–improves, improving, etc.). The basic, stemmed forms of our keywords are presented in Table 1.

Table 1. Goal-related keywords.

Verb Keywords	*Noun Keywords*
make, improve, increase, promote, develop, formulate, prepare, reduce, maintain, administer, guarantee, offer, prolong, endorse, manage, obtain, avoid, block, prevent, achieve, require, lack, ensure, motivate, decrease, reduce, enhance, enable, support, provide, aim	objective, aim, purpose, goal

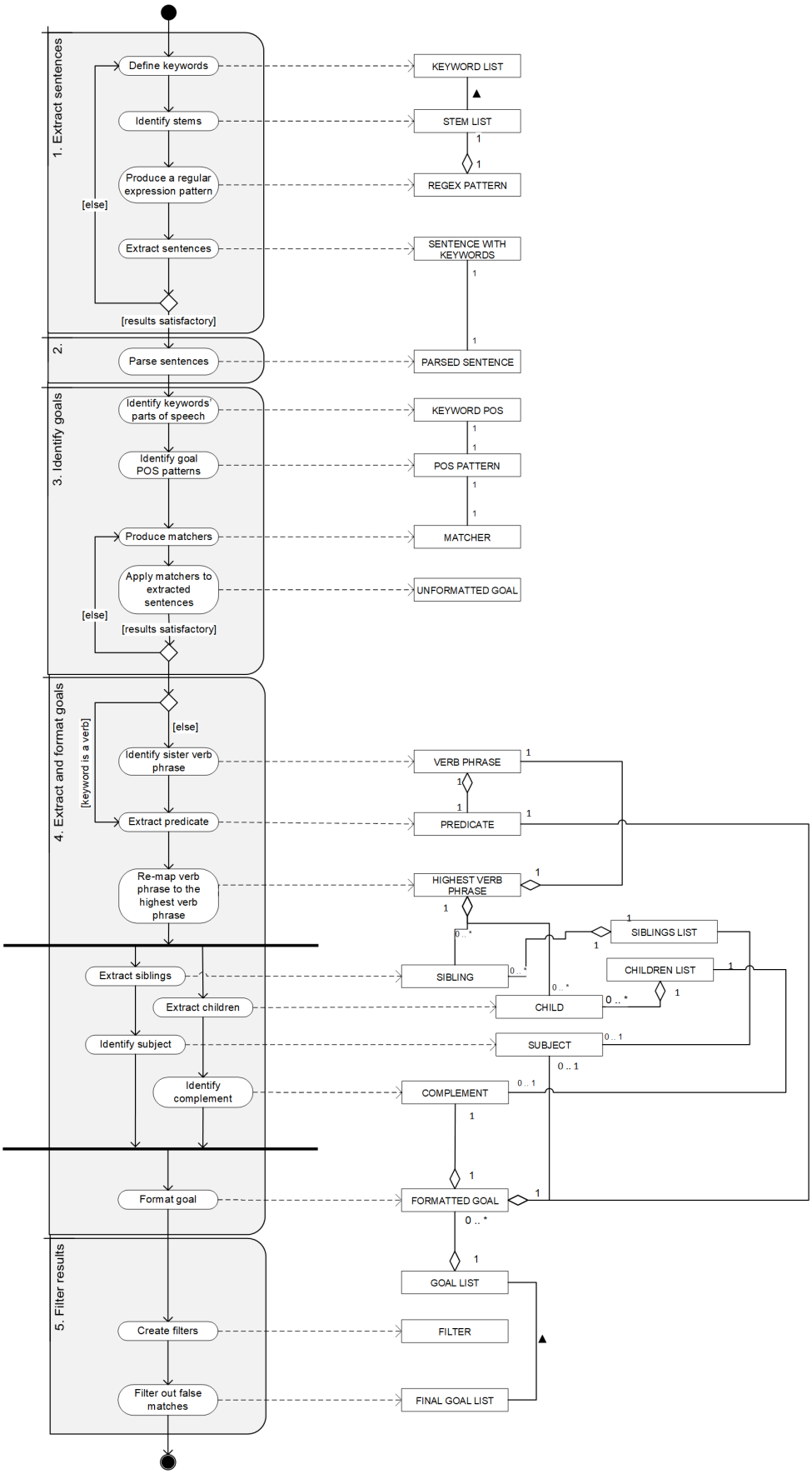

Figure 2. The NLP-Automated Method for Business Understanding (NAMBU) generates business goals from an organization's email conversations.

The sentence extraction component uses regular expressions to identify the sentences, which are suspected to contain business goals, by extracting them in full, when they contain the stem of the keyword. The regular expression is composed of:

- A negated set, repeating 0 or more times, which does not include a full stop, an exclamation point, and a question mark. Essentially, it extracts anything, which is in between these signs and the stem of the keyword;
- A disjunction of all the stems of the previously identified goal-related keywords (e.g., "promot", instead of "promote", to match not only with "promote" or "promotes", but also with "promoting");
- Another set of 0 or more signs, which do not signify the end of the sentence;
- An expression matching one of: a full stop, an exclamation point, or a question mark, which means that the sentence is suspected to end there;
- An optional expression matching up to one letter, followed by another full stop. This expression can repeat up to two times, to account for the use of acronyms.

Essentially, matching against this pattern results in strings, which are limited by a full stop, a question mark, or an exclamation point both at the beginning and at the end, and which contain the stem of the keyword. If the punctuation and the formatting of the text are correct, this regular expression extracts almost exclusively full sentences. One simplification was made here, which concerns the longer acronyms, as well as abbreviations and contractions. These acronyms may sometimes contain more than three full stops (C.O.B.R.A.). The abbreviations and contractions, such as "Prof." or "Revd.", can also unexpectedly end the extraction, before it reaches the end of the sentence. However, with their relatively low prevalence in extracted text snippets, and their use mostly outside of the key parts of the sentences, namely the business goals, we decided to sacrifice these sentence fragments in the interest of lowering the computational load by limiting the number of all sentences subjected to parsing.

2.2. Parse Syntax

The output is passed as a sentence list onto the next component, which makes use of the Stanford CoreNLP Natural Language Processing Toolkit [18]. The annotators in a pipeline are picked through the basic Java properties in a Properties object. This object is also responsible for specifying the order in which the annotators are applied. The object is further passed to a pipeline, and can be applied to the list of sentences extracted in the previous step of the process. The artefact of this project uses the following annotators:

- A Penn Tree Bank tokenizer [18], which tokenizes the text into a sequence of tokens. It has been developed with web text in mind, thus it deals quite well with noise, which can be found in electronic communication between people;
- A sentence splitter, which divides a sequence of tokens into sentences, thus preparing them for further analysis;
- A maximum entropy part-of-speech tagger [24], which labels the tokens with their corresponding part-of-speech tags;
- A parser, which provides a full syntactic analysis of the input text. The phrase structure trees are developed in this stage of analysis [25,26].

Figure 3 illustrates an example sentence after parsing, represented as a phrase structure tree. The sentence was extracted by our method from the Enron corpus due to the inclusion of the keyword "provide": *"We have also committed to provide more segmented information about our business units and how we operate so that analysts have a better understanding of our businesses"*. Figure 3 displays the part-of-speech tags for this example sentence, arranged in a hierarchy. For example, at the top level, the example sentence (S) is constructed by concatenating a noun phrase (NP) with a verb phrase (VP). The NP part is resolved by replacing it with the personal pronoun (PRP) 'we'. The VP is rewritten with the non-3rd person present verb 'have', the adverb (ADVP) phrase 'also', and an embedded VP 'committed [...]'.

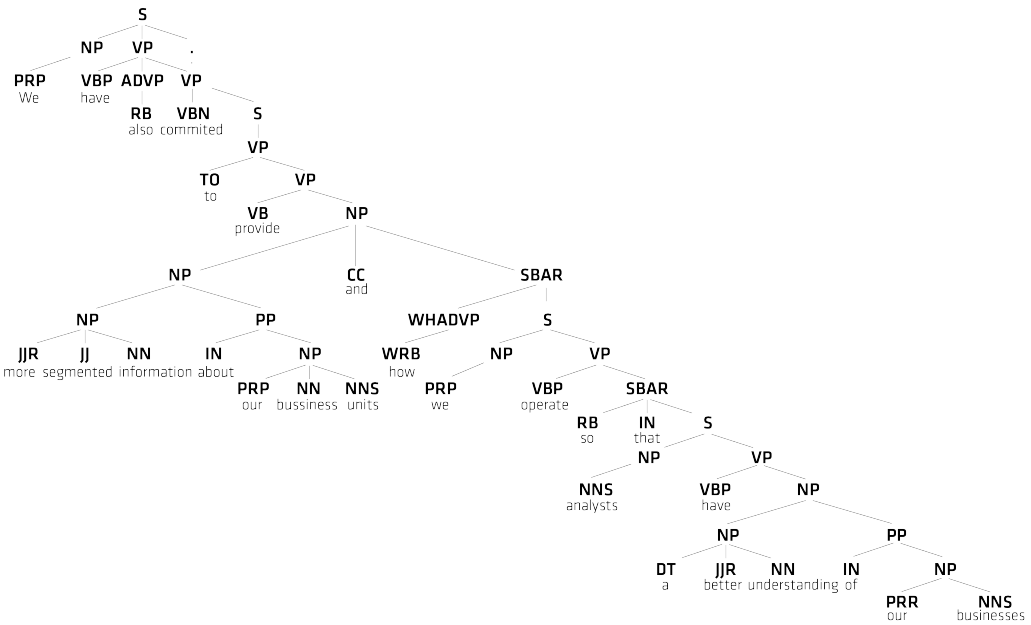

Figure 3. Phrase structure tree created by parsing of the example sentence within the Enron corpus.

2.3. Identify Goals

Tregex [27] is used for extracting information, based on the part-of-speech and chunk tags. It allows for identification and extraction of trees matching specified patterns. For instance "NP < NN $ VP" identifies a noun phrase, which is a parent of a singular form of a noun, while at the same time being a sister to a verb phrase. These patterns can get much more complicated than that, and extract much more specific content.

The extracted goal was defined as one matching the *[(subject) predicate complement]* formula, with the subject optional (extracted only if it exists); the predicate being the verb within the sentence, and the complement referring to the rest of the sub-tree (the object, the prepositional phrase, and/or subordinate conjunction). The algorithm extracts only these triplets, which include keywords. This was done to ensure high precision of the mechanism, instead of focusing on the recall, due to statistical reasons. With a large set of data for analysis, some goals may not be identified, but the majority should be extracted. This provides a significant boost to the usability of the method, by reducing the effort needed for the second-level, manual goal identification. The matchers take the parsed sentences as input, and investigate if the suspected goal is in fact a target for extraction. In the case of the keyword being a verb, the matcher looks for the highest verb phrase VP, dominating a verb in any form (tags: VB, VBZ, VBP, VBD, VBN, VBG), which directly dominates one of the previously defined verb keywords. A similar mechanism is used for the noun keywords, in this case the algorithm looks for the highest noun phrase NP, dominating a singular or plural noun tag (NN, NNS), which directly dominates the keyword noun. The output is composed of the excerpts from the sentences, which match the patterns.

2.4. Extract and Format Goals

The highest subtree matching the noun phrase, or a verb phrase is passed to this component, which does the following: in case of a verb phrase, it extracts the verb-keyword as the predicate, identifies the siblings and the children subtrees, and explores them looking for, respectively, the subject and the complement. In case of input being a noun phrase, it first identifies the sister verb phrase, and then does the same. This mechanism is presented on the NAMBU Process-Deliverable Diagram in Figure 2. The results are formatted as [(subject) predicate complement] triplets and recorded as identified business goals.

2.5. Filter Results

At this stage, a goal list is created, however one more step is necessary to enhance its precision, thus usability. This is the automatic filtering of unwanted false positive results. This component was developed through an analysis of a random sample of 447 e-mails from the Enron corpus, while looking for common ground between the false positive matches, which was not shared with true positive matches. This analysis yielded the following results:

- A high number of analyzed false positives concerns personal information exchange, which is not related to the organization, however gets labeled as a match due to the appearance of keywords within it. A high number of this type of false positives has the participants of the information exchange addressing each other personally, for example, by using the word "you". This is not as prominent within the true positives group, where the e-mail authors rarely address the recipients personally within the same sentence (or–sentence fragment) that contains the business goal;
- Another pattern identified within the false positives group concerns the organization of: lunches, dinners, meetings, conferences, sleeping accommodation and so forth. The language used often contains keywords, such as "provide" or "ensure";
- False positives tend to contain first and/or last names of people, for example, "Provide John with an update". This is not encountered that often within the true positives, however there are instances where an excerpt from a sentence can be classified as a business goal, even while explicitly stating a person's name;

Having these observations, the following filters were applied:

- Any match that personally addresses the recipient of the e-mail is filtered out, thus not being included in the results. This had an extremely positive effect on the precision of the tool, while having an insignificantly negative effect on the recall;
- Any match that includes the words "lunch", "dinner", "meeting", "conference", "accommodation" is filtered out as well. This filtering had a slightly positive effect on precision, while not affecting the recall at all;
- Since filtering first and last names from the results would require a next instance of a higher level natural language processing algorithm, which would in turn need much more computational power and time, while not improving the combined measure of precision and recall significantly, this idea was abandoned;

It is important to note that this step of the method is not as generalizable as the other method steps. Depending on the area of application, whether it is a specific business, or an industry, the keywords for filtering might need to be re-adjusted.

2.6. Evaluation

To evaluate the artefact's effectiveness, we picked another set of e-mails from the Enron corpus, which had not been used during the development. We chose the */maildir/lay-k/sent* folder, which contained 266 e-mails sent by the CEO of Enron and his assistants, as well as quotations of the e-mails he, or his assistants, were responding to. The e-mails were carefully read (by MK), and each of the business goals written down within them was marked as one, and saved. Later, 124 matches were compared to the output of the tool. This was done to determine the precision and recall of the tool, and compile the f-measures.

3. Results

In Information Retrieval with binary classification (either relevant or not relevant), the effectiveness of an algorithm is commonly measured by precision, recall, and the F-measure (also known as F1 score, or F-score), which combines these two scores into one measurement [28].

The evaluation of the artefact on the Enron CEO's sent folder from the Enron corpus resulted in the following values, as shown in Table 2:

Table 2. Results of the application of the artefact to the Enron CEO's *sent* folder.

True Positives	*False Positives*	*False Negatives*	*Precision*	*Recall*	*F-Measure*
74	22	50	0.77	0.597	0.672

The 22 false positives, along with 50 false negatives, were then subjected to further analysis on the basis of comparison to true positives and the syntactic pattern used for extraction.

A correctly identified business goal contains the keyword and matches the syntactic pattern. An example of a correctly identified business goal extracted from the Enron Corpus is *develop advertising and other ancillary revenue streams*. It consists of the keyword *develop*, which dominates a noun phrase advertising and other ancillary revenue streams. This pattern is, however, not infallible, as some false positives, such as *promote the notion of a person's personal or religious beliefs*, are also matched through it. The keyword *promote* dominates the noun phrase *the notion of a person's personal or religious beliefs*. In the strict sense of the word, it is still a goal, however, not a business goal. This is the most common type of occurrence among the false positives, which proves a limitation of the keywords approach. As close as it is to the identification of exclusively business goals, other types of goals can, and will be extracted too, as the fluidity of natural language does not limit particular verbs to relate solely to business situations. Another possible occurrence of a false match, even though it has not been identified within the results, is still a risk nonetheless. Due to the matching and extraction of only fragments of sentences in the final stages of the process, a negated business goal can slip through and be incorrectly identified as a business goal, as a result of the pattern missing the negation. This proves the evaluation of the results by the stakeholders to be still necessary at the end of the process.

As for the false negatives, all of the 50 omissions belong to one of two groups, as shown in Table 3:

Table 3. Examples of reasons for false negatives.

Reason for Omission	*Example*
no keywords	deal with issues that impede, or facilitate, market development (...)
no verb phrase	reduc = e = 20the risk of black outs this summer

The second example actually does contain a verb phrase; however, due to wrong formatting of the text, and the artefacts "=" and "=20" present, the verb phrase is incorrectly annotated as a noun phrase, which causes rejection by the algorithm when compared to the syntactic pattern. While the lack of a verb phrase is difficult to deal with using this method, the other reason for omission, namely the lack of keywords, can be reduced in its impact by further analyses of goal-phrase-patterns and readjustments of the list of keywords, which triggers the sentences to be extracted.

Concerning the applicability of the results to the business understanding phase of CRISP-DM, the matter is more subjective and cannot be assessed quantitatively with full confidence. A broader context is needed for the business goals, as well as the organization's data mining capabilities. Assessing the matter subjectively, however, results in finding the potential for data mining goals, or at least identification of areas for data mining within 62% of the true positive results.

4. Discussion

The results of the application of the NAMBU goal extraction method are quite promising. The main advantage of applying the method is a significant time gain by reducing the vast amounts of text found in 266 e-mails to 96 potential matches, including 74 correct ones. We expect the usefulness of the method to increase in parallel with the size of the dataset. The modular design of the artefact also provides the opportunity for further adjustments of the keywords, as well as the filters, which may be needed after further analyses of formulas

for business goals, or just for the application of the method at a business within a specific industry. At this stage, it is inevitable for some encoded business goals to slip through the algorithm and not be recorded as matches, and the keywords and filters are the main cause of this limitation.

The potential practical implications of the development of this method cannot be understated. Even in case of a business goal not being directly translatable to a data mining goal, thus not working within the business understanding phase of CRISP-DM, the method can be applied within a wider context of the overall knowledge management or even master data management of the enterprise (e.g., [29,30]. Systematic deployment of the artefact at an organization may provide a significant boost to the automation of business goal tracking.

The usage of a relatively dated dataset is still a significant limitation of this research. As this type of data is usually extremely sensitive, we had to conduct our research using the only available resources of this type, that is, the publicly available Enron Corpus. The content of the e-mails within this dataset is quite outdated, when compared to the e-mail information flow in modern organizations. A lack of insider information about the organization is also quite apparent, when it comes to business goals with no context around them. As the manual business goal identification was conducted internally, and there exists no specific formula for a business goal, this identification was performed in a subjective manner, with an inherent possibility of categorization bias. This validity threat of the study should be addressed in future research, preferably using more recent datasets.

Multiple directions for future research can be taken [31]. We believe that more theoretical, linguistic research on a clear definition of a business goal, or a definition of a business goal with data mining capabilities could take place. Undoubtedly, if such a definition existed, it would be more feasibly translatable to the appropriate natural language processing terms, which would make the business goals easier to extract with higher precision and recall results. Considering how beneficial a framework for goal extraction through text analysis could be for organizations, not only for data mining, our first recommendation for future research is the development of a clearer formula for encoding business goals.

Concerning more natural extensions to this study, its precision can be enhanced by the application of sentiment analysis and negation handling to entire sentences containing the suspected business goals. When it comes specifically to data mining, and CRISP-DM, the business goals identified through the tool are not always translatable to data mining goals. Furthermore, the significant recent advances in open information extraction (OIE) based on deep learning techniques [32,33] may now provide the opportunity to pursue a subsymbolic approach to goal identification and extraction. Or better yet, NAMBU's symbolic approach could be combined with a deep learning-based subsymbolic approach in a hybrid goal extraction ensemble method [34]. The recently released Stanza NLP toolkit seems especially promising in this regard and even integrates CoreNLP [35].

A method for the identification of data mining goals within the extracted business goals can be another area of future research on this subject, thus bootstrapping business understanding along with, partially, the data understanding phase of CRISP-DM [36]. Another recommendation of ours is a multiple case-study: deployment of the tool at multiple organizations, using their internal e-mail repositories as datasets. This way, the obtained results can be analyzed, and the natural language processing component, along with the filtering component, can be readjusted according to these results. This would measure the generalizability of this research, and make it possible to enhance it, wherever the room for improvement can be found. Finally, employing a business process analysis methodology such as process mining may prove to be a useful approach to help determine data mining goals through the identification of an organization's bottlenecks and deviations [37].

5. Conclusions

This paper presents a novel approach of applying natural language processing techniques to corporate e-mail repositories to facilitate the identification and formulation of an organization's business goals. The NLP-Automated Method for Business Understanding (NAMBU) artefact of the research provides a recipe for the extraction of the sentences suspected to contain business goals, as well as their further natural language processing analysis and filtering. The results were evaluated on the basis of the application of the artefact to the Enron corpus. It showed the advantages of using natural language processing as a tool for business goal extraction from large datasets. These repositories contain a lot of knowledge about the organization at hand, at the same time being virtually impossible to explore manually. Automating this process removes this barrier, as well as reducing the impact of this privacy-sensitive situation of reading through the employees' correspondences.

As for the applicability of the results, using information extraction techniques for automatic business goal detection can provide a significant boost not only for the definition of explicit business goals in data mining projects, but also for the overall knowledge management of the organization. The results do, however, need to be manually evaluated at the end of the process for quality control purposes, as false positive results are still unavoidable for such a natural language processing task.

Author Contributions: Conceptualization, M.S.; methodology, M.K., V.M. and M.S.; software, M.K.; validation, M.K., M.S. and V.M.; formal analysis, M.K.; investigation, M.K., V.M. and M.S.; writing—original draft preparation, M.K.; writing—review and editing, M.S.; visualization, M.K.; supervision, M.S. and V.M.; project administration, M.K., V.M. and M.S.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: This research reuses the Enron corpus by Klimt and Yang (2004).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chapman, P.; Clinton, J.; Kerber, R.; Khabaza, T.; Reinartz, T.; Shearer, C.; Wirth, R. *CRISP-DM 1.0: Step-by-Step Data Mining Guide*; SPSS Inc.: Chicago, IL, USA, 2000; Volume 9, p. 13.
2. Sharma, S.; Osei-Bryson, K.M.; Kasper, G.M. Evaluation of an integrated Knowledge Discovery and Data Mining process model. *Expert Syst. Appl.* **2012**, *39*, 11335–11348. [CrossRef]
3. Wang, H.; Wang, S. A knowledge management approach to data mining process for business intelligence. *Ind. Manag. Data Syst.* **2008**, *108*, 622–634. [CrossRef]
4. Becher, J.D.; Berkhin, P.; Freeman, E. Automating exploratory data analysis for efficient data mining. In Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Virtual Event, Memphis, TN, USA, 6–10 July 2000; pp. 424–429.
5. Linoff, G.S.; Berry, M.J. *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*; John Wiley & Sons: Hoboken, NJ, USA, 2011.
6. Sharma, S.; Osei-Bryson, K.M. Framework for formal implementation of the business understanding phase of data mining projects. *Expert Syst. Appl.* **2009**, *36*, 4114–4124. [CrossRef]
7. Marbán, O.; Segovia, J.; Menasalvas, E.; Fernández-Baizán, C. Toward data mining engineering: A software engineering approach. *Inf. Syst.* **2009**, *34*, 87–107. [CrossRef]
8. Menger, V.; Spruit, M.; Hagoort, K.; Scheepers, F. Transitioning to a data driven mental health practice: Collaborative expert sessions for knowledge and hypothesis finding. *Comput. Math. Methods Med.* **2016**, *2016*. [CrossRef] [PubMed]
9. Spruit, M.; Meijers, S. The CRISP-DCW Method for Distributed Computing Workflows. In Proceedings of the International Research & Innovation Forum, Geneva , Switzerland, 11–12 February 2019; pp. 325–341.
10. Lefebvre, A.; Spruit, M.; Omta, W. Towards reusability of computational experiments: Capturing and sharing Research Objects from knowledge discovery processes. In Proceedings of the 2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), Lisbon, Portugal, 12–14 November 2015; Volume 1, pp. 456–462.
11. Stai, E.; Kafetzoglou, S.; Tsiropoulou, E.E.; Papavassiliou, S. A holistic approach for personalization, relevance feedback & recommendation in enriched multimedia content. *Multimed. Tools Appl.* **2018**, *77*, 283–326.
12. Liddy, E.D. Natural Language Processing 2001. Available online: https://surface.syr.edu/istpub/63/ (accessed on 17 September 2021).

13. Campbell, C.S.; Maglio, P.P.; Cozzi, A.; Dom, B. Expertise identification using email communications. In Proceedings of the Twelfth International Conference on Information and Knowledge Management, New Orleans, LA, USA, 3–8 November 2003; pp. 528–531.
14. Merali, Y.; Davies, J. Knowledge capture and utilization in virtual communities. In Proceedings of the 1st International Conference on Knowledge Capture, Victoria, BC, Canada, 22–23 October 2001; pp. 92–99.
15. Grobelnik, M.; Mladenic, D.; Fortuna, B. Semantic technology for capturing communication inside an organization. *IEEE Internet Comput.* **2009**, *13*, 59–67. [CrossRef]
16. Spruit, M.; Jagesar, R. Power to the People! In Proceedings of the International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, Porto, Portugal, 9–11 November 2016; pp. 400–406.
17. Spruit, M.; Lytras, M. *Applied Data Science in Patient-Centric Healthcare: Adaptive Analytic Systems for Empowering Physicians and Patients*; Elsevier: Amsterdam, The Netherlands, 2018.
18. Manning, C.D.; Surdeanu, M.; Bauer, J.; Finkel, J.R.; Bethard, S.; McClosky, D. The Stanford CoreNLP natural language processing toolkit. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Baltimore, MD, USA, 22–27 June 2014; pp. 55–60.
19. Klimt, B.; Yang, Y. The enron corpus: A new dataset for email classification research. In Proceedings of the European Conference on Machine Learning, Pisa, Italy, 20–24 September 2004; pp. 217–226.
20. van de Weerd, I.; Brinkkemper, S. Meta-modeling for situational analysis and design methods. In *Handbook of Research on Modern Systems Analysis and Design Technologies and Applications*; IGI Global: Hershey, PA, USA, 2009; pp. 35–54.
21. Van Lamsweerde, A. *Requirements Engineering: From System Goals to UML Models to Software*; John Wiley & Sons: Chichester, UK, 2009; Volume 10.
22. Casagrande, E.; Woldeamlak, S.; Woon, W.L.; Zeineldin, H.H.; Svetinovic, D. NLP-KAOS for systems goal elicitation: Smart metering system case study. *IEEE Trans. Softw. Eng.* **2014**, *40*, 941–956. [CrossRef]
23. Lezcano R, L.A.; Guzmán L, J.A.; Gómez A, S.A. Extraction of goals and their classification in the KAOS model using natural language processing. *Ingeniare. Rev. Chil. Ing.* **2015**, *23*, 59–66. [CrossRef]
24. Toutanova, K.; Klein, D.; Manning, C.D.; Singer, Y. Feature-rich part-of-speech tagging with a cyclic dependency network. In Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, Edmonton, AB, Canada, 27 May–1 June 2003; pp. 252–259.
25. Klein, D.; Manning, C.D. Accurate unlexicalized parsing. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, Sapporo, Japan, 7–12 July 2003; pp. 423–430.
26. De Marneffe, M.C.; MacCartney, B.; Manning, C.D. Generating typed dependency parses from phrase structure parses. *Lrec* **2006**, *6*, 449–454.
27. Levy, R.; Andrew, G. Tregex and Tsurgeon: Tools for querying and manipulating tree data structures. *LREC* **2006**, 2231–2234. Available online: https://nlp.stanford.edu/pubs/levy_andrew_lrec2006.pdf (accessed on 17 September 2021).
28. Goutte, C.; Gaussier, E. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In Proceedings of the European Conference on Information Retrieval, Santiago de Compostela, Spain, 21–23 March 2005; pp. 345–359.
29. Wijaya, S.; Spruit, M.R.; Scheper, W.J. Webstrategy formulation: Benefiting from web 2.0 concepts to deliver business values. In *Web 2.0*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 1–30.
30. Spruit, M.; Pietzka, K. MD3M: The master data management maturity model. *Comput. Hum. Behav.* **2015**, *51*, 1068–1076. [CrossRef]
31. Kais, M. Bootstrapping the CRISP-DM Process. Master's Thesis, Utrecht University, Utrecht, The Netherlands, 2017.
32. Sarhan, I.; Spruit, M. Can We Survive without Labelled Data in NLP? Transfer Learning for Open Information Extraction. *Appl. Sci.* **2020**, *10*, 5758. [CrossRef]
33. Huang, L. Cold-Start Universal Information Extraction. Ph.D. Thesis, University of Illinois at Urbana-Champaign, Champaign, IL, USA, 2020.
34. Alam, M.; Groth, P.; Hitzler, P.; Paulheim, H.; Sack, H.; Tresp, V. CSSA'20: Workshop on Combining Symbolic and Sub-Symbolic Methods and their Applications. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management, Online, 19–23 October 2020; pp. 3523–3524.
35. Qi, P.; Zhang, Y.; Zhang, Y.; Bolton, J.; Manning, C.D. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Online, 5–10 July 2020.
36. Huber, S.; Wiemer, H.; Schneider, D.; Ihlenfeldt, S. DMME: Data mining methodology for engineering applications—A holistic extension to the CRISP-DM model. *Procedia Cirp* **2019**, *79*, 403–408. [CrossRef]
37. Wu, Q.; He, Z.; Wang, H.; Wen, L.; Yu, T. A business process analysis methodology based on process mining for complaint handling service processes. *Appl. Sci.* **2019**, *9*, 3313. [CrossRef]